Data-Driven Keyword Marketing Optimization for LabelMaster

In the Spring 2025 CSP 572 Data Science Practicum course under Dr. Yong Zheng, I undertook a project titled **"Keyword Marketing Optimization for LabelMaster."** The primary goal was to develop a data-driven system to optimize keyword marketing strategies, with two major focuses: **analyzing organic versus paid search performance** to identify budget-saving opportunities and **predicting keyword profitability** to prioritize future marketing investments. I consolidated and merged real-world datasets from **Google Ads**, **Bing Paid Search**, **SEO platforms**, **sales transactions**, and **GA4 order medium data**, spanning bi-monthly periods from **February 2022 to January 2025**. Using key metrics like search volume, impressions, paid clicks, costs, CTR, and revenue, we performed extensive feature engineering, creating composite scores for organic (**Organic_Score**) and paid (**Paid_Score**) performance by log-transforming skewed features, scaling them using Min-Max normalization, and averaging the results to build holistic indicators of keyword strength.

I further segmented keywords into three natural tiers (Low, Mid, High) for both organic and paid channels using **KMeans clustering** to enable dynamic, data-driven thresholding instead of static cutoffs. Based on these scores, I trained a **Random Forest Classifier** to predict business labels that categorized keywords into strategic action groups — such as cutting spend, monitoring, or investing further — achieving strong classification stability across folds. Additionally, I developed a **custom Gym-like environment** and trained a **Deep Q-Network (DQN)** agent using **TensorFlow/Keras** to simulate real-world budget reallocation between Paid and SEO, successfully training the agent to maximize cumulative marketing ROI.

For the **ROI-based keyword profitability prediction**, I designed two labeling strategies based on ROI thresholds (≥1.5 for a conservative model and ≥1.2 for a more inclusive model). I then trained and compared the following four machine learning models:

- **Logistic Regression**: Chosen as the **baseline** due to its simplicity and interpretability. It provided a basic benchmark but suffered from relatively lower accuracy and F1 scores, especially on imbalanced keyword classes.

- **Random Forest Classifier**: Selected for its **robustness and resistance to overfitting**. It improved performance by learning from multiple decision trees but was still not optimal in maximizing recall and F1-scores for critical profitable keywords.
- **Gradient Boosting Classifier**: A **sequential boosting method** that showed further improvements by correcting errors made by previous weak learners. It achieved higher accuracy and more balanced precision-recall metrics.
- **XGBoost Classifier**: An **optimized and scalable version of gradient boosting**. XGBoost incorporates regularization, advanced tree pruning, and efficient parallelization, making it highly powerful for structured datasets. It outperformed all other models, achieving a **97.1% overall accuracy**, a **0.93 recall** for "Keep Paying" profitable keywords, and a **0.96 F1-score**, indicating extremely high precision and minimal false positives.

Given XGBoost's superior performance across all key metrics — including its ability to capture nearly all profitable keywords accurately while minimizing classification errors — it was selected as the **final model** for deployment. It was integrated into a **Streamlit web application** to ensure usability for non-technical marketing teams. The application supports manual keyword entries, bulk CSV uploads, and generates downloadable prediction files, allowing easy and efficient use without requiring technical expertise.

In conclusion, the project successfully delivered a comprehensive, scalable system that enables LabelMaster to optimize advertising spend, improve marketing ROI, identify SEO weaknesses, and automate keyword profitability forecasting. The combination of traditional supervised models, advanced ensemble methods, and reinforcement learning techniques positioned the solution as a future-ready, intelligent marketing optimization platform. Planned future enhancements include expanding to campaign- and adgroup-level optimization and introducing time series-based ROI forecasting to capture seasonal dynamics.